PersonaCLR: Evaluation Model for Persona Characteristics via Contrastive Learning of Linguistic Style Representation

Michimasa Inaba
The University of Electro-Communications
1-5-1, Chofugaoka, Chofu, Tokyo, Japan
m-inaba@uec.ac.jp

Abstract

Persona-aware dialogue systems can improve the consistency of the system's responses, users' trust and user enjoyment. Filtering nonpersona-like utterances is important for constructing persona-aware dialogue systems. This paper presents the PersonaCLR model for capturing a given utterance's intensity of persona characteristics. We trained the model with contrastive learning based on the sameness of the utterances' speaker. Contrastive learning enables PersonaCLR to evaluate the persona characteristics of a given utterance, even if the target persona is not included in training data. For training and evaluating our model, we also constructed a new dataset of 2,155 character utterances from 100 Japanese online novels. Experimental results indicated that our model outperforms existing methods and a strong baseline using a large language model. Our source code, pre-trained model, and dataset are available at https://github.com/1never/PersonaCLR.

1 Introduction

Persona-aware dialogue systems can improve the consistency of the system's responses (Li et al., 2016), users' trust in the system (Higashinaka et al., 2018), and user enjoyment (Miyazaki et al., 2016).

In constructing persona-aware dialogue systems, automatic estimation of persona characteristics' intensity is important in two ways. First, if we can detect low-intensity utterances of persona characteristics, inappropriate system responses can be prevented. Second, the automatic measure helps construct the persona's sample utterance set. Two methods for constructing persona-aware dialogue systems include the following: (1) using persona descriptions (Zhang et al., 2018; Song et al., 2019; Majumder et al., 2020; Kim et al., 2020; Tang et al., 2023) and (2) sample utterances (Higashinaka et al., 2018; Mitsuda et al., 2022; Han et al., 2022). In the method using samples, if we can filter out samples not matching the persona, the system's performance will improve. This paper presents an evaluation model for Persona characteristics via Contrastive learning of Linguistic style Representation (PersonaCLR), which can measure a given utterance's intensity of the target persona's characteristics. In this paper, the term persona indicates both real-life individuals and fictional characters. PersonaCLR receives the evaluation target's utterance and a target persona's set of sample utterances and then returns a score indicating the target persona characteristics' intensity. The model is trained by contrastive learning based on the sameness of the utterances' speaker. Contrastive learning enables PersonaCLR to evaluate the persona characteristics of a given utterance, even if the target persona is not included in training data.

To the best of our knowledge, there are no public datasets for the training and evaluation of models to assess the intensity of persona characteristics in utterances. We constructed and published two datasets: the Naro Utterance dataset (NaroU), containing 2,155 characters' utterances from 100 Japanese online novels, and an evaluation dataset based on dialogue scenarios between a user and a character in NaroU. We use the dataset to train and evaluate PersonaCLR in the experiments. The creation of these datasets contributes to the advancement of research on dialogue systems that mimic fictional characters. Additionally, this dataset can be utilized for speaker identification tasks (He et al., 2013; Muzny et al., 2017; Yu et al., 2022) and for building persona-aware dialogue systems using sample utterances (Han et al., 2022). The evaluation dataset for this task was also constructed and published.

This study contributes the following: (1) a new model for assessing the intensity of target persona characteristics in a given utterance that does not require retraining or fine-tuning, even if the persona is not included in the training data; (2) a new

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 674–685
September 18–20, 2024. ©2024 Association for Computational Linguistics

open dataset including over 2000 character utterances from 100 Japanese online novels and annotated human-character dialogue scenarios; and (3) a demonstration of the effectiveness of our model using a comparison with existing methods and a strong baseline involving ChatGPT.

2 Related Work

2.1 Persona Characteristics Evaluation

Persona-based dialogue models have been actively studied (Song et al., 2019; Majumder et al., 2020; Kim et al., 2020, 2022) since the release of the PERSONA-CHAT dataset (Zhang et al., 2018). These models receive a dialogue context and a few sentences of persona description (e.g., “I have two dogs.") and then include the description’s content as much as possible in generated responses. Several proposed evaluation metrics for these models evaluate generated utterances according to how much of a given persona description’s content is included. Both Persona F1 (Jiang et al., 2020a) and Persona coverage (Jiang et al., 2020a) are metrics that utilize nonstop words common between a given persona description and an utterance. Persona accuracy (Zheng et al., 2020), which predicts whether a given persona description is exhibited in generated utterances, is computed by feeding generated responses into a binary classifier and obtaining classification accuracy.

This study focuses on persona-aware dialogue systems that mimic a fictional character rather than on systems based on persona descriptions. Since defining such personas with only a few descriptive sentences is difficult, several methods have been proposed to construct such persona-aware dialogue systems using a few samples (Han et al., 2022) or manually collected responses (Higashinaka et al., 2018; Mitsuda et al., 2022). For the same reasons as above, systems’ evaluation by methods based on the given persona description’s content is difficult. Therefore, Persona Speaker Probability (PSProb) (Miyazaki et al., 2021) and Persona Term Salience (PTSal) (Miyazaki et al., 2021) have been proposed as evaluation metrics for dialogue systems’ utterances that mimic a fictional character’s persona.

2.2 Contrastive Learning

In computer vision, contrastive unsupervised representation learning has been proposed, and performance in object detection and image segmentation has significantly improved (He et al., 2020). The key idea is that this type of learning minimizes the distance between feature representations of different views of the same image and maximizes between-feature representations of views of different images (Chen et al., 2020). Contrastive learning has also been applied to natural language processing, and various models for learning sentence representations have been proposed (Fang et al., 2020; Chen et al., 2020; Giorgi et al., 2021).

Particularly relevant to our study is supervised contrastive learning (Khosla et al., 2020; Gunel et al., 2021; Zhang et al., 2022), which constructs positive and negative pairs by leveraging ground truth labels. Inspired by supervised contrastive learning, we constructed contrastive learning pairs based on the sameness of the utterances’ speaker.

2.3 Novel Dataset and Speaker Identification

Several corpora with speaker annotations based on novels have been constructed for several languages: the Columbia Quoted Speech Attribution Corpus (Elson and McKeown, 2010), P&P (He et al., 2013), QuoteLi3(Muzny et al., 2017), and RiQuA (Papay and Padó, 2020) are English corpora; WP (Chen et al., 2019, 2021), JINYONG (Jia et al., 2020) and CSI (Yu et al., 2022) are Chinese; and RWG (Brunner, 2013) is German. In these corpora, speaker annotations were performed for a few of the novels (the highest was 18 in CSI). Thus, the diversity of worldviews and characters are limited. We annotated 100 online novels written in Japanese and constructed and released a new dataset.

Existing corpora have been mainly constructed for speaker identification (SI), that is, to identify the corresponding speaker(s) for each utterance in novels (He et al., 2013; Muzny et al., 2017; Yu et al., 2022; Chen et al., 2023). In SI, an utterance and its surrounding context are given, and, using the context, SI models determine the utterance’s speaker. Our task can be regarded as predicting a given utterance’s speaker, but because no context is given in our task, we cannot apply existing SI methods.

3 PersonaCLR

Our task is to estimate the intensity of the characteristics of a target persona c within a given utterance x. The existing SoTA model, PSProb (Miyazaki et al., 2021), is based on a multi-class classification model classifying which character uttered the given input utterance. Therefore, when evaluat-

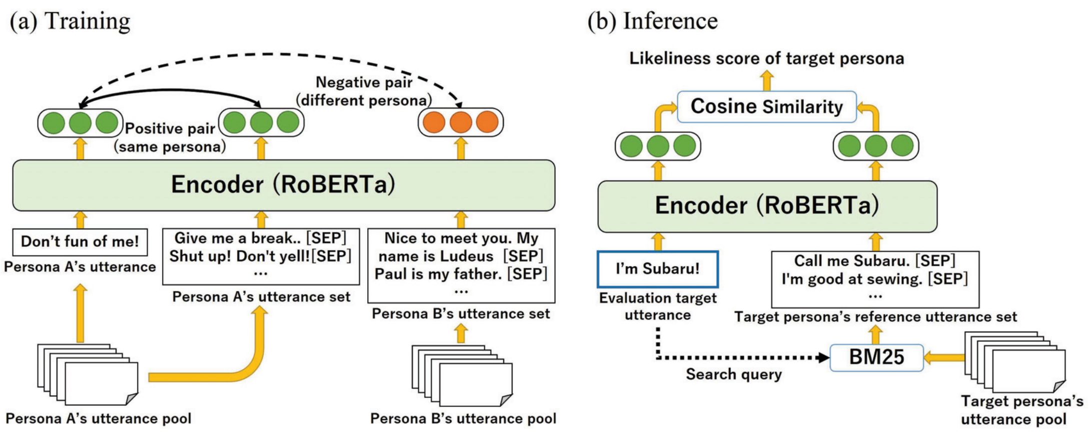

Figure 1: (a) Training with contrastive learning in PersonaCLR. An utterance and an utterance set were sampled from the utterance pool and encoded by the Transformer encoder RoBERTa. Pairs of encoded utterance and utterance sets are used as positive pairs for contrastive learning if they are sampled from the same persona's utterance pool and as negative pairs otherwise. (b) Inference in PersonaCLR. The reference utterance set is constructed from the target character's utterance pool using BM25. The likeliness score is obtained as the cosine similarity between encoded vectors from the target utterance and the utterance set.

ing the persona characteristics of a new character, PSProb must collect not only the reference utterance sets for the evaluation of the target persona as well as the utterance sets of non-target personas, then training the model from scratch.

We propose PersonaCLR, which does not require the retraining and utterance sets for non-target personas. PersonaCLR uses contrastive learning to distinguish whether a given utterance and a set of utterances come from the same persona. Note that PersonaCLR does not evaluate the similarity of any given utterance to utterances in the training data; rather, it observes the similarity between the given utterance and a reference set of utterances. Therefore, our model can evaluate utterances of characters that are not included in the training data without requiring retraining, and it requires only a small number (≥ 20) of references.

We define two embedding vectors of the same speaker's utterances as positive pairs and two vectors by different speakers as negative pairs for contrastive learning. However, because utterances do not necessarily reflect persona characteristics, one of the pair's embedding vectors is obtained from a set of utterances rather than from a single utterance. In contrastive learning, the model distinguishes whether an utterance and a set of utterances are from the same persona.

3.1 Training and Inference

An overview of PersonaCLR's training is shown in Figure 1 (a). Let $x^a = \{x^a_i\}^n_{i=1}$, be the utterances pool by a speaker a. By sampling x^a, we obtain an utterance x^a_k and reference utterance set $x^{a+} = \{x^a_j\}^m_{j=1}$. We use the pair of the utterance x^a_k and reference utterance set x^{a+} as positive pair for contrastive learning. On the other hand, we use the pair of the utterance x^a_k, and utterance set x^{b+}, sampled from speaker b's utterance pool x^b as negative pair.

Each utterance and reference utterance set is encoded by the Transformer encoder RoBERTa (Liu et al., 2020). Before encoding, the utterance sets are concatenated with a separator token $[SEP]$ to form a single sequence. With RoBERTa, we obtain the embedding vectors $\mathbf{h}^a_k$ and $\mathbf{h}^{a+}$ corresponding to x^a_k and x^{a+}.

The loss function using these embedded vectors is defined as follows:

$$l^a_k = -log \frac{e^{sim(\mathbf{h}^a_k, \mathbf{h}^{a+})/\tau}}{\sum_{i=1}^{N} e^{sim(\mathbf{h}^a_k, \mathbf{h}^{s_i+})/\tau}} \quad (1)$$

where N is the batch size and s_i is the speaker of the ith utterance set in the batch. The τ is the temperature hyperparameter and $sim(h_1, h_2)$ is the cosine similarity $\frac{\mathbf{h_1}^\top \mathbf{h_2}}{||\mathbf{h_1}|| \cdot ||\mathbf{h_2}||}$.

In the inference phase, the persona-likeness score of a given utterance is the cosine similarity between the utterance's embedding vectors and the target persona's reference set in Figure 1 (b).

3.2 Related Utterance Sampling

A reference utterance set should contain sufficient information to evaluate the utterance's persona characteristics. For example, if the target utterance contains a habit or terminology unique to the persona, for appropriate evaluation, the reference utterance set should also include it.

To this end, we ranked the utterance pool x^a using BM25 (Robertson et al., 1995). We used evaluation target utterance x^a_k as a query and obtain the top m utterances as the reference set x^{a+} in the inference phase (see Figure 1 (b)). In the experiment, all utterances in the training data of the NaroU (see Section 4.5) were used to calculate parameters for then calculating BM25 (average number of words per utterance and IDF). In the inference phase, we use the reference utterances set using BM25. In the training phase, we simultaneously use the utterance set using BM25 and the randomly sampled set as training data to ensure robustness.

4 Experiment

Figure 2 shows a summary of experimental procedure used to evaluate the effectiveness of PersonaCLR. In this experiment, we constructed and used two types of datasets: NaroU and the evaluation dataset. The NaroU dataset consists of utterances from novels. However, there is the concern that, compared to utterances in novels, utterances in a dialogue between a user and a persona-aware dialogue system differ in length and tendency. To address this concern, we created the evaluation dataset.

4.1 Naro Utterance Dataset (NaroU)

For training models to assess the intensity of persona characteristics in utterances, we constructed NaroU, a dataset of utterances in novels annotated with speaker attributions. This dataset was constructed by annotating 100 novels in “*Shosetsuka ni Naro*,"[1] a Japanese novel self-publishing website[2]. Most of the website’s novels are divided into episodes of 2000 to 5000 Japanese characters each, and we annotated each novel’s first ten episodes. We recruited annotators via the crowdsourcing website CrowdWorks[3] and instructed them to extract segments of utterances in the novel and assign speaker names. We instructed annotators to annotate the speaker’s real name if it was given in the novel or otherwise, a nickname or pronoun. One annotator performed annotation per each novel. We paid them 600 JPY (approximately 4 USD) per episode. Table 1 shows the statistics of this dataset.

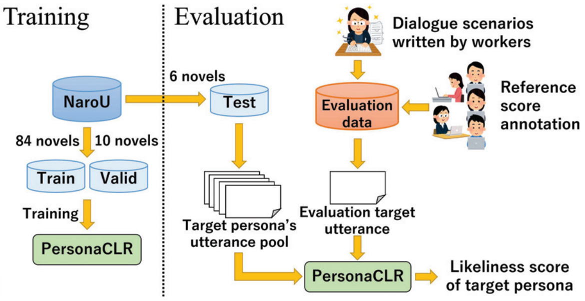
Training
Evaluation
Dialogue scenarios written by workers
6 novels
NaroU
Test
Evaluation data
Reference score annotation
84 novels
10 novels
Train
Valid
Training
PersonaCLR
Target persona's utterance pool
Evaluation target utterance
PersonaCLR
Likeliness score of target persona

Figure 2: Experimental overview. Two datasets were used, NaroU (blue) and evaluation data (red). NaroU is divided into train, valid, and test datasets so that novels and characters do not overlap among the three. The six novels and their characters in the test data are shown in Table 2. Evaluation data is based on dialogue scenarios between a user and a target character created by crowd workers. All character utterances in the dialogue scenarios are annotated with reference scores of persona characteristics. During the evaluation phase, PersonaCLR is given an evaluation target utterance in dialogue scenarios and the corresponding persona’s utterance pool; it then outputs an estimated score.

Novels	100
Episodes	1000
Characters	2,155
Utterances	38,297
Words in utterances	620,820

Table 1: Statistics of the NaroU

To confirm annotations’ consistency, we conducted an experiment in which 50 episodes of five novels were individually annotated by two annotators. Experimental results showed that the perfect agreement rate was 88.4%.

NaroU is divided into train, valid, and test datasets (84, 10, and 6 novels, respectively). The train and valid datasets are used to train the model, and the test dataset is used as the utterance pool (see Figure 1 (b)). Train, valid, and test datasets contained no overlap of novels and characters.

4.2 Evaluation Data Collection

Utterances from novels are not suitable for assessing PersonaCLR’s ability to estimate the intensity of a persona in a system’s utterances. For this experiment, therefore, we created dialogue scenarios in which a user interacts with a character and then used these scenarios’ utterances.

[1] https://syosetu.com/
[2] “*Shosetuka ni Naro*" means “*Let’s become a novelist.*"
[3] https://crowdworks.jp/

Ncode	Character
n6316bn	Rimuru (75), Veldora (63)
n9669bk	Rudeus (264), Roxy (140) Paul (105), Sylphiette (65), Zenith (60)
n2267be	Subaru (220)
n4830bu	Myne (187), Tuuli (83), Effa (63)
n3191eh	Leon (353), Luxion (86), Angelica (76), Olivia (67)
n5040ce	Catarina (190), Keith (52)

Table 2: List of target characters used for collecting evaluation data. Ncode is a unique ID assigned to each novel submitted to *Shosestuka ni Naro*. Numbers in brackets indicates the number of unique utterances in the test dataset of NaroU used as the utterance pool of the target character. The utterances in the novels and characters shown in this table were not included in the training data for PersonaCLR.

4.2.1 Dialogue Scenario

To collect the evaluation dataset, we recruited workers via CrowdWorks. They created dialogue scenario between a specified character and a user. For a situation in which a user is talking with a character, each scenario was individually created by one worker. We paid them 250 JPY (approximately 1.5 USD) per created dialogue.

We selected characters as evaluation target personas based on the following two conditions:(1) novels within the top 100 in cumulative ranking calculated on the number of bookmarks and reviews on *Shosetsuka ni Naro* and (2) novels developed as media mixes in both manga and anime as of January 2023. We selected 17 regular characters with 50 or more unique utterances from novels that satisfied the two conditions above. The list of characters is shown in Table 2, in which the ncode is an identifier uniquely assigned to each novel. We accessed the novels at https://ncode.syosetu.com/(ncode)/.

We prepared 13 general topics for dialogue (e.g., hobbies, travel, and family), selected by the workers. Each scenario consisted of 10 utterances, spoken alternately by the user and the character. We recruited only workers who had watched at least ten episodes of a novel's anime, read 10 episodes of the novel, or two books of the novel's manga containing the character. Through the procedure described above, we collected 20 dialogues (100 utterances) per character, for a total of 1,700 utterances.

4.2.2 Reference Score Annotation

To obtain reference scores of persona characteristics in the evaluation dataset, crowd workers annotated the collected utterances. Workers were paid 70 JPY (approximately 0.4 USD) per ten utterances. The definition and assignment procedure of the reference score complied with the previous study (Miyazaki et al., 2021). We only extracted the target character's utterances from the created scenarios. We asked workers to answer with “yes" or “no" whether the character was likely to have said each utterance. The workers evaluated them only by observing the utterances, without considering the context. Five people evaluated each utterance, and the number of “yes" answers was used as the reference score.

Characters	17
Unique utterances	1,700
Words per utterance	31.56
Reference scores	
- Score 5	967 (43.1%)
- Score 4	422 (18.8%)
- Score 3	168 (7.5%)
- Score 2	120 (5.3%)
- Score 1	153 (6.8%)
- Score 0	414 (18.4%)
Total	2,244

Table 3: Statistics of evaluation data

Because the number of low scores was small for only utterances created as target characters, we also annotated utterances created as other characters. The previous study (Miyazaki et al., 2021) used 500 utterances and scores (100 utterances of target characters and 400 utterances of non-target characters). However, this setting was far from reality, with, in some cases, a score of 0 accounting for 60% of the total. Therefore, we annotated 100 utterances of the target character and two randomly sampled utterances of non-target characters for each character, for a total of 132 utterances ($= 100 + 16$ characters $\times 2$). finally, we obtained 2,244 evaluation scores ($= 132 \times 17$ characters). Table 3 displays evaluation data statistics, and Table 3 shows examples of utterances and scores.

4.3 Comparative Methods

4.3.1 Persona Speaker Probability (PSProb)

PSProb (Miyazaki et al., 2021) is a previous SoTA method that uses multi-class classification with logistic regression. A set of utterances for each persona was prepared as training data, and logistic regression was trained so each utterances could be classified as being from any of the personas. At the time of inference, the probability that the target utterance is by the target persona is calculated by

	Utterance	Score
Myne	本さえあれば何もいらないと思っているよ。 (As long as I have books, I do not need anything.)	5
User	やっぱり読書をしていたら時間を忘れちゃう？ (Do you lose track of time when you read books?)	-
Myne	そんなことはしょっちゅうあった。 (That happened to me often.)	4
User	今度おすすめの本を紹介してくれる？ (Can you recommend a book for me sometime?)	-
Myne	勿論！あなたにぴったりの本を紹介するね！ (Sure! I'll introduce you to the book that's right for you!)	5

Table 4: Example of a dialogue scenario and reference scores. The dialogue topic is reading, and Myne is a character from *Ascendance of a Bookworm* (ncode: n4830bu). The scores are the number of people of the five annotators who judged the utterance as Myne-like.

logistic regression, then used as a score.

In PSProb's original configuration, training data were equalized for each character, so our experiment also used this configuration. As Table 2 shows, the smallest number of utterances among all characters was 52 for Keith. Therefore, for each of the 17 characters, we used 52 utterances, 50 for training data and 2 for development data, for 882 utterances in total[4].

4.3.2 Persona Term Salience (PTSal)

PTSal (Miyazaki et al., 2021) is a method for assigning scores to terms in a given utterance. The method is based on TF-IDF and assigns higher scores to terms more frequently used by the target character and less by others. The term scores' average is used as the estimated utterance score.

4.3.3 ChatGPT

In recent years, the performance of large language models (LLMs), such as GPT-3 (Brown et al., 2020), has improved significantly on few-shot settings that use only a few examples. We used ChatGPT (gpt-3.5-turbo) (OpenAI, 2022) as a strong baseline in this experiment. ChatGPT outputs the target utterance's likeliness score as an integer value from 0 to 5. For the utterance list, we used the top m utterances ranked by BM25 using the target utterances as a query, as with PersonaCLR. The examples were randomly selected from six utterances with a score of 0 to 5, one by one, from the target character's evaluation data. Since examples' order affects results in the few-shot prompting (Gao et al., 2021; Jiang et al., 2020b; Liu et al., 2021), we shuffled the six examples' order. The parameters given to the ChatGPT API were set to default settings except for temperature, which was set to 0.0 to generate deterministically. Appendix A.2 shows an example of the actual prompt and the hyperparameters of ChatGPT.

4.4 BERTScore

BERTScore (Zhang et al., 2019) calculates similarity between texts by using vector representations obtained from pre-trained BERT. We calculated BERTScore between all pairs of the target utterance and reference utterances; the maximum BERTscore was used as the target utterance's score. As reference utterances, we used the target character's utterances in the NaroU (Table 2)

4.4.1 MaxBLEU

MaxBLEU(Xu et al., 2018) is the maximum BLEU score between all pairs of the target utterance and reference utterances. We used SacreBLEU (Post, 2018) to compute the BLEU score.

4.4.2 Persona-F1 (P-F1)

Rather than reference utterance-based, P-F1 (Jiang et al., 2020a) is a persona description-based evaluation measure that evaluates how well persona characteristics are expressed in an utterance. The higher the overlap between the non-stop word in the persona description and the utterance, the higher the P-F1 score.

4.5 Implementation Details

We trained PersonaCLR using data from 94 out of the 100 novels in the NaroU dataset. We excluded the six novels shown in Table 2 to prevent any overlap between the characters in the training data in the NaroU and the evaluation data described in Section 5.1. We used 84 of the 94 novels as training data and ten as development data. We used Japanese RoBERTa$_{large}$[5] for PersonaCLR and BERTScore. We used the size of reference utterance set m to 20 in PersonaCLR and ChatGPT.

For PersonaCLR and PSProb, we conducted hyperparameter optimization. To find the optimal hyperparameters of PersonaCLR, a grid search was performed with temperature τ as $\{0.01, 0.05, 0.1\}$, batch size as $\{16, 32, 64\}$, warmup steps as

[4] Previous study (Miyazaki et al., 2021) used 55 utterances for each character.

[5] https://huggingface.co/nlp-waseda/roberta-large-japanese-with-auto-jumanpp

Character	PersonaCLR	PSProb	PTSal	ChatGPT	BERTScore	MaxBLEU	P-F1
Rimuru	0.201	0.015	0.005	0.124	**0.261**	0.099	-0.046
Veldora	**0.614**	0.478	0.386	0.340	0.337	0.450	0.067
Rudeus	**0.663**	0.334	0.426	0.314	0.369	0.392	0.098
Roxy	**0.644**	0.594	0.327	0.372	0.376	0.362	0.283
Sylphiette	**0.696**	0.550	0.540	0.345	0.594	0.240	0.309
Paul	**0.598**	0.311	0.290	0.023	0.288	0.311	0.191
Zenith	**0.527**	0.323	0.148	0.285	0.262	0.084	0.186
Subaru	**0.585**	0.447	0.236	0.222	0.120	0.165	0.218
Myne	**0.415**	0.147	0.150	0.145	0.181	0.042	0.181
Tuuli	**0.481**	0.332	0.308	0.401	0.220	0.202	0.143
Effa	**0.453**	0.295	0.236	0.351	0.207	0.068	0.117
Leon	**0.372**	0.273	0.197	0.120	0.129	0.148	0.245
Olivia	**0.726**	0.457	0.393	0.379	0.468	0.358	0.296
Angelica	**0.518**	0.290	0.374	0.116	0.311	0.172	0.179
Luxion	**0.641**	0.560	0.517	0.349	0.546	0.486	0.323
Catarina	**0.464**	0.328	0.174	0.277	0.293	0.254	0.144
Keith	**0.603**	0.476	0.471	0.420	0.387	0.209	0.297
Average	**0.541**	0.365	0.305	0.270	0.315	0.238	0.190

Table 5: Spearman's rank correlation coefficients (r_s) between the reference and estimated scores.

Character	PersonaCLR	PSProb	PTSal	ChatGPT	BERTScore	MaxBLEU	P-F1
Rimuru	0.395	0.271	0.218	**0.440**	0.391	0.241	0.406
Veldora	**0.887**	0.630	0.417	0.544	0.431	0.633	0.582
Rudeus	**0.783**	0.541	0.605	0.461	0.494	0.531	0.471
Roxy	**0.697**	0.737	0.430	0.469	0.404	0.324	0.574
Sylphiette	**0.808**	0.557	0.579	0.453	0.552	0.485	0.543
Paul	**0.786**	0.609	0.514	0.270	0.362	0.466	0.581
Zenith	**0.748**	0.440	0.271	0.355	0.344	0.201	0.545
Subaru	**0.866**	0.556	0.347	0.374	0.341	0.328	0.438
Myne	**0.483**	0.246	0.229	0.325	0.224	0.212	0.529
Tuuli	**0.738**	0.355	0.296	0.477	0.285	0.344	0.587
Effa	**0.656**	0.412	0.302	0.608	0.326	0.205	0.601
Leon	**0.665**	0.404	0.359	0.261	0.267	0.272	0.482
Olivia	**0.908**	0.690	0.673	0.511	0.584	0.627	0.616
Angelica	**0.529**	0.339	0.348	0.288	0.294	0.468	0.585
Luxion	**0.758**	0.576	0.598	0.510	0.543	0.666	0.666
Catarina	**0.722**	0.582	0.487	0.637	0.501	0.560	0.590
Keith	**0.770**	0.498	0.510	0.545	0.448	0.479	0.634
Average	**0.718**	0.497	0.423	0.443	0.399	0.414	0.555

Table 6: AUPR for inappropriate utterance filtering

$\{100, 300, 500\}$ and learning rate as $\{1e^{-4}$, $5e^{-5}$, $1e^{-5}, 5e^{-6}, 1e^{-6}\}$, respectively. As a result, we set the temperature τ to 0.05, batch size to 64, warmup step to 300, and learning rate to $1e^{-5}$. During training, the loss of development data was calculated every 100 steps, and the model with the lowest loss was used for evaluation. For PSProb, the hyperparameter C, the inverse of regularization strength, was grid-searched on a logarithmic scale from 0.01 to 100, and C was set to 100.

4.6 Evaluation Indices

We used two indices to examine PersonaCLR's performance and this experiment's comparative methods: Spearman's rank correlation r_s and the area under the precision-recall curve (AUPR). We used Spearman's rank correlation r_s to verify that PersonaCLR scores correlated with the human ratings and AUPR to evaluate PersonaCLR's performance in filtering inappropriate utterances. In calculating AUPR, we regarded utterances with a reference score of 0 or 1 as the detection target.

4.7 Results

Experimental results based on Spearman's rank correlation r_s between the manually assigned reference and estimated scores are shown in Table 5. For PersonaCLR and PSProb whose results depend on the random seed, training was performed three times with different seeds, and average values are shown in Table 5. Our PersonaCLR overperforms all other metrics, including ChatGPT. In PersonaCLR, 15 out of 17 characters showed a moderate correlation or higher (> 0.4), and seven characters showed a strong correlation (> 0.6). PSProb showed the proposed method's second-best performance. On more than half of the characters, PTSal showed inferior correlations to PSProb. ChatGPT and BERTScore results were uncorrelated (< 0.2) for some characters, although in some cases, correlation exceeded the results of PSProb. ChatGPT and BERTScore showed higher performance than MaxBLEU, which also used only the target character's reference utterances. However, ChatGPT and BERTScore were inferior to PersonaCLR and PSProb, which used utterances of several characters in training or calculating scores, thus suggesting that leveraging differences between personas is important in this task. PersonaCLR can learn this difference efficiently through contrastive learning, resulting in high performance. Overall, MaxBLEU, and P-F1 showed low performance, although correlations were observed for some characters.

Results of AUPR for inappropriate utterance filtering are shown in Table 6. PersonaCLR showed the best performance for 16 of 17 characters. One major difference from the results in Table 5 is that P-F1, which included character names and terms in its persona description, performed relatively better. In this experiment, most inappropriate utterances were created by non-target characters from other novels. Therefore, P-F1 effectively filtered out utterances that did not contain specific names or terms. ChatGPT showed relatively high performance for some characters, however, it was inferior to PersonaCLR, PSProb, and P-F1. Our results thus confirm PersonaCLR's effectiveness.

Character	PersonaCLR	w/o BM25	w/ Single Ref.
Rimuru	**0.201**	0.178	0.119
Veldora	**0.614**	0.612	0.580
Rudeus	**0.663**	0.542	0.498
Roxy	**0.644**	0.465	0.534
Sylphiette	**0.696**	0.565	0.616
Paul	0.598	**0.694**	0.495
Zenith	**0.527**	**0.527**	0.470
Subaru	**0.585**	0.529	0.466
Myne	**0.415**	0.367	0.262
Tuuli	0.481	**0.563**	0.462
Effa	**0.453**	0.354	0.334
Leon	0.372	**0.418**	0.310
Olivia	**0.726**	0.621	0.696
Angelica	**0.518**	0.504	0.491
Luxion	**0.641**	0.546	0.609
Catarina	**0.464**	0.447	0.355
Keith	**0.603**	0.382	0.542
Average	**0.541**	0.489	0.461

Table 7: Spearman's rank correlation coefficients (r_s) for ablation models

Character	PersonaCLR	w/o BM25	w/ Single Ref.
Rimuru	0.395	**0.454**	0.325
Veldora	**0.887**	0.884	0.851
Rudeus	**0.783**	0.719	0.711
Roxy	0.697	**0.746**	0.551
Sylphiette	**0.808**	0.670	0.681
Paul	0.786	**0.808**	0.639
Zenith	**0.748**	**0.748**	0.628
Subaru	**0.866**	0.844	0.636
Myne	**0.483**	0.452	0.452
Tuuli	0.738	0.770	**0.780**
Effa	**0.656**	0.599	0.557
Leon	**0.665**	0.615	0.414
Olivia	**0.908**	0.813	0.866
Angelica	**0.529**	0.459	0.411
Luxion	0.758	0.683	**0.775**
Catarina	**0.722**	0.692	0.604
Keith	**0.770**	0.581	0.650
Average	**0.718**	0.679	0.620

Table 8: AUPR for ablation models

5 Ablation Study

We conducted experiments using ablation models. The following two models were compared: A model that randomly samples from a pool of utterances instead of using BM25 to construct a set of reference utterances (w/o BM25), and a model that uses a single utterance as a reference that is the most similar to the target utterance by BM25 instead of the utterance set (w/ Single Ref.).

The results for each character in the ablation models using Spearman's rank correlation coefficient are shown in Table 7 and those using AUPR are shown in 8. PersonaCLR shows the best performance for 14 of 17 characters in rank correlation coefficient, and 12 characters in AUPR. We also found that w/o BM25 outperformed PersonaCLR on several characters. This suggests that BM25 may have constructed an inappropriate reference utterance set for evaluating a given target utterance. Although we used the traditional ranking method BM25 in this paper, performance could be improved by improving the method of constructing reference utterances. With Single Ref., only two characters outperformed PersonaCLR in AUPR and zero in the correlation coefficient. These results indicate that employing a set of utterances rather than just a single utterance was important for appropriate evaluation.

The ablation study reconfirms PersonaCLR's effectiveness.

6 Visualization

The embedding vector $\mathbf{h}$ of the utterance obtained by the Transformer encoder in PersonaCLR reflects persona characteristics, and the same speaker's utterances are closely placed in the vector space. We

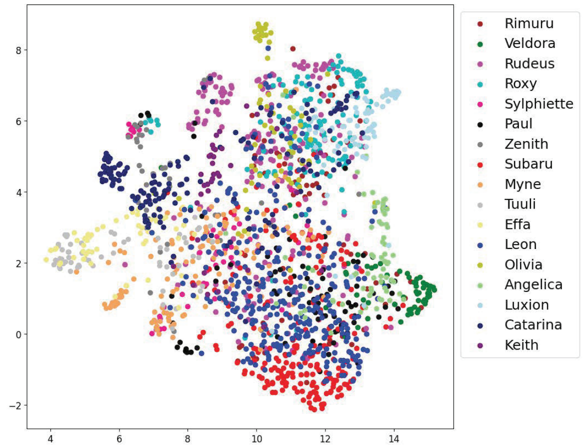

Figure 3: Utterance embedding visualization

visualized the vectors to confirm what speaker features were emphasized in the embedding process.

Figure 3 shows the embedding results of the 17 characters, that is, all utterances encoded by PersonaCLR and dimension reduction by Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). We found that the manner of speaking determines the position. Characters who use polite language (e.g., Rudeus, Roxy, Olivia, Luxion, Keith) are placed in the upper area, those who do not use polite language (e.g., Paul, Subaru, Leon) in the lower, and those who use female language (e.g., Myne, Tuuli, Effa, Catarina) are placed on the left. We can also observe a tendency for characters in the same novel to be placed close. For example, Myne, Tuuli, and Effa, as well as Catarina and Keith, are located near each other due to common use in their utterances of novel-specific terms and character names. In contrast, Leon and Luxion, who is from the same novel, are positioned far apart, indicating that they are embedded with more emphasis on the manner of speaking than on being from the same novel.

7 Conclusion

We proposed a novel model for evaluating a given utterance's intensity of persona characteristics and constructed the Naro Utterance dataset (NaroU) for training our model. The proposed model employs contrastive learning, and experimental results show that our model outperforms existing methods.

Future work includes constructing persona-aware dialogue systems by applying PersonaCLR and evaluating its performance experimentally. We also plan to extend PersonaCLR to be able to evaluate on a context-response basis rather than an utterance basis. This extension is expected to further improve the response performance of the system.

Acknowledgments

This research was supported by the NEDO project "Development of Interactive Story-Type Contents Creation Framework."

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Annelen Brunner. 2013. Automatic recognition of speech, thought, and writing representation in german narrative texts. *Literary and linguistic computing*, 28(4):563–575.

Jia-Xiang Chen, Zhen-Hua Ling, and Li-Rong Dai. 2019. A Chinese dataset for identifying speakers in novels. In *INTERSPEECH*, pages 1561–1565. Graz, Austria.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Yue Chen, Tianwei He, Hongbin Zhou, Jia-Chen Gu, Heng Lu, and Zhen-Hua Ling. 2023. Symbolization, prompt, and classification: A framework for implicit speaker identification in novels. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3455–3467.

Yue Chen, Zhen-Hua Ling, and Qing-Feng Liu. 2021. A neural-network-based approach to identifying speakers in novels. In *Interspeech*, pages 4114–4118.

David Elson and Kathleen McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 1013–1019.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages "3816–3830".

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2021. Supervised contrastive learning for pretrained language model fine-tuning. *Ninth International Conference on Learning Representation, ICLR 2021*.

Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5114–5132.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Ryuichiro Higashinaka, Masahiro Mizukami, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Junji Tomita. 2018. Role play-based question-answering by real users for building chatbots with consistent personalities. In *Proceedings of the 19th annual sigdial meeting on discourse and dialogue*, pages 264–272.

Yuxiang Jia, Huayi Dou, Shuai Cao, and Hongying Zan. 2020. Speaker identification and its application to social network construction for Chinese novels. *International Journal of Asian Language Processing*, 30(04):2050018.

Bin Jiang, Wanyue Zhou, Jingxu Yang, Chao Yang, Shihan Wang, and Liang Pang. 2020a. PEDNet: A persona enhanced dual alternating learning network for conversational response generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4089–4099.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020b. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will i sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916.

Minju Kim, Beong-woo Kwak, Youngwook Kim, Hong-in Lee, Seung-won Hwang, and Jinyoung Yeo. 2022. Dual task framework for improving persona-grounded dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10912–10920.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach. In *Eighth International Conference on Learning Representations, ICLR 2020*. The International Conference on Learning Representations.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29).

Koh Mitsuda, Ryuichiro Higashinaka, Hiroaki Sugiyama, Masahiro Mizukami, Tetsuya Kinebuchi, Ryuta Nakamura, Noritake Adachi, and Hidetoshi Kawabata. 2022. Fine-tuning a pre-trained transformer-based encoder-decoder model with user-generated question-answer pairs to realize character-like chatbots. In *Conversational AI for Natural Human-Centric Interaction: 12th International Workshop on Spoken Dialogue System Technology, IWSDS 2021, Singapore*, pages 277–290. Springer.

Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, and Yoshihiro Matsuo. 2016. Towards an entertaining natural language generation system: linguistic peculiarities of Japanese fictional characters. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 319–328.

Chiaki Miyazaki, Saya Kanno, Makoto Yoda, Junya Ono, and Hiromi Wakaki. 2021. Fundamental exploration of evaluation metrics for persona characteristics of text utterances. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 178–189.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470.

OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt.

Sean Papay and Sebastian Padó. 2020. RiQuA: A corpus of rich quotation annotation for English literary text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages "835–841".

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. *In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5190—-5196.

Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5456–5468.

Zhen Xu, Nan Jiang, Bingquan Liu, Wenge Rong, Bowen Wu, Baoxun Wang, Zhuoran Wang, and Xiaolong Wang. 2018. Lsdscc: a large scale domain-specific conversational corpus for response generation with diversity oriented evaluation metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2070–2080.

Dian Yu, Ben Zhou, and Dong Yu. 2022. End-to-end Chinese speaker identification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2274–2285.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Zhenyu Zhang, Yuming Zhao, Meng Chen, and Xiaodong He. 2022. Label anchored contrastive learning for language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1437–1449.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Mao Xiaoxi. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9693–9700.

A Details of Comparative Methods

A.1 Persona Term Salience (PTSal)

PTSal (Miyazaki et al., 2021) is a method for assigning scores to terms in a given utterance. The term scores' average is used as the estimated utterance score for evaluation. PTSal is obtained by the following equation.

$$PTSal(t,p) = UttFreq(t,p) \cdot SpkrRarity(t) \tag{2}$$

$$UttFreq(t,p) = \frac{n(t,p)}{m(p)} \tag{3}$$

$$SpkrRarity(t) = log\frac{|P|}{s(t)} \tag{4}$$

where $n(t,p)$ is the number of utterances with term t in the monologue of persona p and $m(p)$ is the total number of utterances in the monologue of persona p. $s(t)$ is the number of personas that used term t, and $|P|$ is the total number of personas. The $UttFreq(t,p)$ becomes larger the more the target persona uses the term t, and $SpkrRarity(t)$ is larger if a small number of personas other than the target persona uses the term t. In the experiment, we calculated $SpkrRarity(t)$ using all the utterances in the NaroU.

== Task ==
Based on examples of a character's utterances below, assign a rating from 0 to 5 to indicate the probability that the utterance was spoken by the character.

== Character's utterance examples ==
- Come on, it's a tie-in. The knife and machete are sinking first, but what about you?
- Oh, I was a shut-in!
- Do not think less of me just because I'm a shut-in. My grip strength was over seventy kilograms. I can bench press up to 80 kilos!
...
(The rest is omitted. 20 utterances in total)

== Rating examples ==
Utterance: Uh... I'm not good at horror...
Rating: 2/5

Utterance: I do not watch many movies, but the only movie I watched recently was "One Piece."
Rating: 1/5

Utterance: Okay, I'll buy it for you! I'll get it for you, just wait there.
Rating: 5/5

Utterance: My hobby is to learn all kinds of skills! Sewing, embroidery, figure skating, magic tricks... you name it!
Rating: 3/5

Utterance: We get into trouble from time to time, but we live well together.
Rating: 0/5

Utterance: Well, that settles it then. Yeah, it looks good on you.
Rating: 4/5

Utterance: Seriously, seriously, I'm soooo happy, looking forward to it!
Rating:

Table 9: Prompt format example for ChatGPT (originally written in Japanese)

A.2 ChatGPT

An example of a prompt used in ChatGPT, which was used as a comparison method, is shown in Table 9. The target character to be evaluated in this example is Subaru from *Re: Life in a Different World from Zero*. The last utterance (bold font) is the evaluation target utterance. ChatGPT generates the score of the utterance after "Rating:." The parameters given to the ChatGPT API were set to default settings except for temperature, which was set to 0.0 to generate deterministically.

A.3 Persona-F1 (P-F1)

Rather than reference utterance-based, P-F1 (Jiang et al., 2020a) is a persona description-based evaluation measure that evaluates how well persona characteristics are expressed in an utterance. P-F1 is calculated as follows:

$$\text{PersonaF1} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \quad (5)$$

$$\text{Recall} = \frac{\max_{i \in [1,L]} |W_{Y \cap d_i}|}{|W_{d_i}|} \quad (6)$$

$$\text{Precision} = \frac{\max_{i \in [1,L]} |W_{Y \cap d_i}|}{|W_Y|} \quad (7)$$

where W_Y is a set of non-stop words in utterance Y and W_{d_i} is a set of non-stop words in sentence d_i in the persona description.

We used character descriptions in *Japanese Wikipedia* as the persona description. Examples of persona descriptions are shown in Table 10. The stop word was determined using the ja-stopword-remover library (version 0.2.4)[6].

本作の主人公。 (The protagonist of this work.)
4月1日生まれ。 (Born on April 1.)
黒の短髪、平凡な顔立ち、筋肉質のがっちりした体格の持ち主である少年。 (He is a teenager with short black hair, an ordinary face, and a stocky, muscular build.)
一般的な日本人よりも速く、目つきの悪さ（三白眼）が特徴である。 (He is faster on his feet than the average Japanese, and he has bad eyesight (*sanpaku* eyes).)
年齢は17歳（開始時点）。 (He is 17 years old (at the beginning of the story).) ...

Table 10: Persona description example of Subaru in *Re: Life in a Different World from Zero* (ncode: n2267be)

[6] https://github.com/Pickerdot/ja_stopword_remover